PATENT ABSTRACTS OF JAPAN

BC

(11)Publication number : 11-143899
(43)Date of publication of application : 28.05.1999

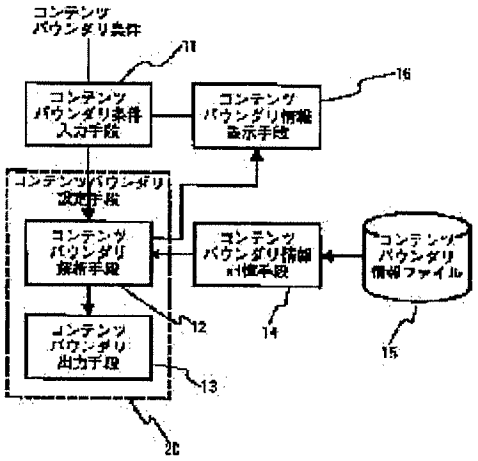
(51)Int.Cl. G06F 17/30
G06F 17/27
G06F 17/21

(21)Application number : 09-307726
(22)Date of filing : 10.11.1997

(71)Applicant : SEIKO EPSON CORP
(72)Inventor : TANAKA TOSHIO

(54) DEVICE AND METHOD FOR REGISTER-DOCUMENT PROCESSING AND STORAGE MEDIUM FOR STORING PROGRAM FOR PROCESSING REGISTERED DOCUMENT

(57)Abstract:
PROBLEM TO BE SOLVED: To make it possible to extract contents of an appropriate size in accordance with a processing.
SOLUTION: This processing device which processes a document that has registered contents boundary information for indicating a boundary of unity of document contents, has at least a contents boundary input means 11 that can input contents boundary condition for taking out contents in accordance with the processing, and a contents boundary set means 20 that sets a contents boundary position on the basis of a contents boundary information set for a registered document by contents boundary condition inputted from this contents boundary condition input means 11; and takes out the contents of the registered document on the basis of the contents boundary position information outputted from the contents boundary set means 20.



(19)日本国特許庁（ＪＰ）　(12)**公開特許公報**（Ａ）　(11)特許出願公開番号

特開平11－143899

(43)公開日　平成11年(1999)５月28日

(51)Int.Cl.[6]		識別記号	ＦＩ		
Ｇ０６Ｆ	17/30		Ｃ０６Ｆ	15/40	３７０Ａ
	17/27			15/20	５５０Ｅ
	17/21				５７０Ｒ
					５８０Ｌ
					５９０Ｅ

審査請求　未請求　請求項の数９　ＯＬ　（全 16 頁）　最終頁に続く

(21)出願番号　特願平9－307726

(22)出願日　平成９年(1997)11月10日

(71)出願人　000002369
セイコーエプソン株式会社
東京都新宿区西新宿２丁目４番１号

(72)発明者　田中　敏雄
長野県諏訪市大和３丁目３番５号　セイコーエプソン株式会社内

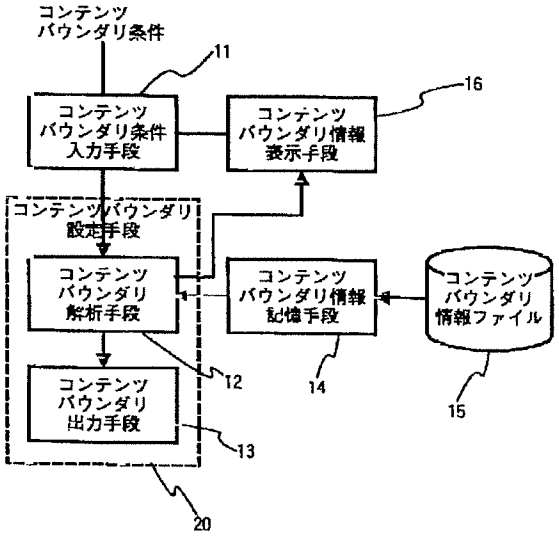
(74)代理人　弁理士　鈴木　喜三郎　（外２名）

(54)【発明の名称】　登録文書処理装置及び方法並びに登録文書を処理するための処理プログラムを記憶した記憶媒体

(57)【要約】

【課題】　文書間の差分を取ったり、検索を行ったりというような処理を行う段階でコンテンツを取り出そうとすると、処理量が多く処理に時間を要する問題がある。

【解決手段】　文書内容のまとまりの境界を示すコンテンツバウンダリ情報を有して登録されている文書を処理するためのものであって、処理に応じたコンテンツを取り出すためのコンテンツバウンダリ条件を入力可能なコンテンツバウンダリ入力手段１１と、このコンテンツバウンダリ条件入力手段１１から入力されたコンテンツバウンダリ条件により、前記登録文書に対して設定されているコンテンツバウンダリ情報に基づくコンテンツバウンダリ位置を設定するコンテンツバウンダリ設定手段２０とを少なくとも有し、コンテンツバウンダリ設定手段２０により出力されたコンテンツバウンダリ位置情報に基づいて登録文書のコンテンツを取り出す。



【特許請求の範囲】
【請求項１】　文書内容のまとまりの境界を示すコンテンツバウンダリ情報が設定されて登録されている文書（登録文書という）を処理するための登録文書処理装置において、
処理に応じたコンテンツを取り出すためのコンテンツバウンダリ条件を入力可能なコンテンツバウンダリ条件入力手段と、
このコンテンツバウンダリ条件入力手段から入力されたコンテンツバウンダリ条件に基づいて、前記登録文書に対して設定されたコンテンツバウンダリ情報を出力し、そのコンテンツバウンダリ情報に対応するコンテンツバウンダリ位置を設定するコンテンツバウンダリ設定手段と、
を少なくとも有することを特徴とする登録文書処理装置。
【請求項２】　前記コンテンツバウンダリ設定手段は、コンテンツバウンダリ条件を受けると、このコンテンツバウンダリ条件と、前記登録文書に対して設定されたそれぞれのコンテンツバウンダリ情報とを比較し、前記コンテンツバウンダリ条件に適合するコンテンツバウンダリ情報を得て、このコンテンツバウンダリ情報に基づくコンテンツバウンダリ位置を設定することを特徴とする請求項１記載の登録文書処理装置。
【請求項３】　文書内容のまとまりの境界を示すコンテンツバウンダリ情報が設定されて登録されている文書（登録文書という）を処理するための登録文書処理装置において、
処理に応じたコンテンツを取り出すために必要なコンテンツバウンダリ条件知識を予め蓄えたコンテンツバウンダリ条件知識ファイルと、
このコンテンツバウンダリ条件知識ファイルの内容に基づいて、前記登録文書に対して設定されたコンテンツバウンダリ情報を処理に応じて出力し、そのコンテンツバウンダリ情報に対応するコンテンツバウンダリ位置を設定するコンテンツバウンダリ設定手段と、
を少なくとも有することを特徴とする登録文書処理装置。
【請求項４】　文書内容のまとまりの境界を示すコンテンツバウンダリ情報が設定されて登録されている文書（登録文書という）を処理するための登録文書処理方法において、
処理に応じたコンテンツを取り出すためのコンテンツバウンダリ条件が入力されると、そのコンテンツバウンダリ条件を受け付け、そのコンテンツバウンダリ条件に基づいて、前記登録文書に対して設定されているコンテンツバウンダリ情報を出力し、このコンテンツバウンダリ情報に対応するコンテンツバウンダリ位置を設定する処理を登録文書処理に含むことを特徴とする登録文書処理方法。
【請求項５】　前記コンテンツバウンダリ条件に基づいて、前記登録文書に対して設定されたコンテンツバウンダリ情報を出力する処理は、入力されたコンテンツバウンダリ条件と、前記登録文書に対して設定されたそれぞれのコンテンツバウンダリ情報とを比較し、前記コンテンツバウンダリ条件に適合するコンテンツバウンダリ情報を得て、このコンテンツバウンダリ情報に基づくコンテンツバウンダリ位置を設定することを特徴とする請求項４記載の登録文書処理方法。
【請求項６】　文書内容のまとまりの境界を示すコンテンツバウンダリ情報が設定されて登録されている文書（登録文書という）を処理するための登録文書処理方法において、
処理に応じたコンテンツを取り出すために必要なコンテンツバウンダリ条件知識を格納したコンテンツバウンダリ条件知識ファイルを有し、このコンテンツバウンダリ条件知識ファイル内容に基づいて、前記登録文書に対して設定されたコンテンツバウンダリ情報を処理に応じて出力し、そのコンテンツバウンダリ情報に対応するコンテンツバウンダリ位置を設定する処理を登録文書処理に含むことを特徴とする登録文書処理方法。
【請求項７】　文書内容のまとまりの境界を示すコンテンツバウンダリ情報が設定されて登録されている文書（登録文書という）を処理するための処理プログラムを記憶した記憶媒体であって、その処理プログラムは、処理に応じたコンテンツを取り出すためのコンテンツバウンダリ条件が入力されると、そのコンテンツバウンダリ条件を受け付け、そのコンテンツバウンダリ条件に基づいて、前記登録文書に対して設定されたコンテンツバウンダリ情報を出力し、このコンテンツバウンダリ情報に対応するコンテンツバウンダリ位置を設定する処理を含むことを特徴とする登録文書を処理するための処理プログラムを記憶した記憶媒体。
【請求項８】　前記コンテンツバウンダリ条件に基づいて、前記登録文書に対して設定されたコンテンツバウンダリ情報を出力する処理は、入力されたコンテンツバウンダリ条件と、前記登録文書に設定されたコンテンツバウンダリ情報とを比較し、前記コンテンツバウンダリ条件に適合するコンテンツバウンダリ情報を得て、このコンテンツバウンダリ情報に基づくコンテンツバウンダリ位置を設定することを特徴とする請求項７記載の登録文書を処理するための処理プログラムを記憶した記憶媒体。
【請求項９】　文書内容のまとまりの境界を示すコンテンツバウンダリ情報が設定されて登録されている文書（登録文書という）を処理するための処理プログラムを記憶した記憶媒体であって、その処理プログラムは、登録文書に対する処理に応じて、処理に応じたコンテンツを取り出すために必要なコンテンツバウンダリ条件をコンテンツバウンダリ条件知識ファイルから取り出し、そ

のコンテンツバウンダリ条件に基づいて、前記登録文書に対して設定されたコンテンツバウンダリ情報を出力し、そのコンテンツバウンダリ情報に対応するコンテンツバウンダリ位置を設定する処理を登録文書処理に含むことを特徴とする登録文書を処理するための処理プログラムを記憶した記憶媒体。

【発明の詳細な説明】

【０００１】

【発明の属する技術分野】本発明は、コンテンツバウンダリ情報が設定されて登録されている文書（登録文書）を用いて、検索処理や複数の文書間の差分を取ったりする処理を行うための登録文書処理装置及び方法並びに登録文書を処理するための処理プログラムを記憶した記憶媒体に関する。

【０００２】

【従来の技術】２つの文書の差分を取ったり、ある文書の一部を抽出したり、文書内に書かれていることを検索したりするというように、文書に対しては様々な処理がなされる。

【０００３】このような処理を行う場合、文書を段落など、文書の内容のまとまり（コンテンツと呼ぶ）ごとにそのまとまりの境界を示すコンテンツバウンダリを検出して、そのコンテンツバウンダリにより得られるコンテンツごとに処理を行う方法が従来より用いられている。

【０００４】従来では、前述したような２つの文書の差分を取ったりする様々な処理を行う際に、コンテンツバウンダリを検出して、コンテンツを抽出するのが一般的である。

【０００５】なお、検索処理を行う場合は、文書中のキーワードや文字列とその位置をインデクス情報として、文書を登録する時に作成しておき、その情報を用いて検索を行うことで検索処理を高速化することが従来より行われている。このような検索処理にあっても、文書をコンテンツに分割する処理は、検索するときに行われるのが普通である。

【０００６】このように、従来では、差分を取ったり、検索したりする処理を行う際、これらの処理を行うに必要なコンテンツの抽出は、それらの処理を行うときになされるのが普通である。

【０００７】しかし、差分を取ったり、検索したりする処理を行う際に、コンテンツを抽出するための処理（コンテンツバウンダリ検出も含めた処理）を行うと、差分を取ったり、検索したりする処理を行う前に、まず、コンテンツを抽出するための処理を行う必要があるため、処理量が多くなり、処理速度の低下を招くことにもなる。つまり、既に登録されている文書に対し、文書間の差分を取ったり、検索したりする処理を行う場合、これらの処理を行うたびに、その都度、文書の構造解析を行って、コンテンツを抽出する処理を行う必要がある。

【０００８】一方、特開平８－２７２８２２の「文書登録装置および文書検索装置」には、文書を登録する際に、文書を所定のブロック（コンテンツと同意のものと考えられるので以下ではコンテンツと表現する）単位に分割し、コンテンツを識別する識別子を付与し、コンテンツ単位からキーワードを抽出し、識別子をキーワードと対応付けしてキーワードのインデクス情報を作成するというような内容が示されている。

【０００９】

【発明が解決しようとする課題】前述の特開平８－２７２８２２（従来技術という）は、文書を登録する際に、文書を所定のコンテンツ単位に分割して登録することが示されている。しかしながら、この従来技術は、登録時に文書を所定のコンテンツに分割してしまうので、検索などの処理時には、そのコンテンツ単位での処理を行うしかなく、ユーザの必要とする大きさのコンテンツを取り出したり、処理内容に応じて、コンテンツを適切な大きさに変更することができないという問題がある。

【００１０】また、この従来技術におけるコンテンツは、検索を行うためのキーワードが文書中のどこに存在するかを表すために用いられるものであり、文書間の差分を取ったり、特定のコンテンツを抽出したりといった様々なコンテンツ処理を行うためのものではないため、これらの処理には不向きである。

【００１１】そこで本発明は、文書を登録する際に、コンテンツの境界を示すコンテンツバウンダリ情報を求め、そのコンテンツバウンダリ情報を有して登録された文書を用いて、検索や文書間の差分を取るといった様々な処理を行い、しかも、既に設定されているコンテンツバウンダリ情報を様々な処理に応じて柔軟に選択可能とすることで、処理に応じた適切な大きさのコンテンツを抽出することを可能とすることを目的としている。

【００１２】

【課題を解決するための手段】前述の目的を達成するために、本発明の請求項１に記載された登録文書処理装置の発明は、文書内容のまとまりの境界を示すコンテンツバウンダリ情報が設定されて登録されている文書（登録文書）を処理するための登録文書処理装置において、処理に応じたコンテンツを取り出すためのコンテンツバウンダリ条件を入力可能なコンテンツバウンダリ条件入力手段と、このコンテンツバウンダリ条件入力手段から入力されたコンテンツバウンダリ条件に基づいて、前記登録文書に対して設定されたコンテンツバウンダリ情報を出力し、そのコンテンツバウンダリ情報に対応するコンテンツバウンダリ位置を設定するコンテンツバウンダリ設定手段とを少なくとも有した構成としている。

【００１３】そして、前記コンテンツバウンダリ設定手段は、コンテンツバウンダリ条件を受けると、このコンテンツバウンダリ条件と、前記登録文書に対して設定されたそれぞれのコンテンツバウンダリ情報とを比較し、前記コンテンツバウンダリ条件に適合するコンテンツバ

ウンダリ情報を得て、このコンテンツバウンダリ情報に基づくコンテンツバウンダリ位置を設定するようにしている。

【００１４】また、請求項３に記載された登録文書処理装置の発明は、文書内容のまとまりの境界を示すコンテンツバウンダリ情報が設定されて登録されている文書（登録文書）を処理するための登録文書処理装置において、処理に応じたコンテンツを取り出すために必要なコンテンツバウンダリ条件知識を予め蓄えたコンテンツバウンダリ条件知識ファイルと、このコンテンツバウンダリ条件知識ファイルの内容に基づいて、前記登録文書に対して設定されたコンテンツバウンダリ情報を処理に応じて出力し、そのコンテンツバウンダリ情報に対応するコンテンツバウンダリ位置を設定するコンテンツバウンダリ設定手段とを少なくとも有した構成としている。

【００１５】また、本発明の請求項４に記載された登録文書処理方法の発明は、文書内容のまとまりの境界を示すコンテンツバウンダリ情報が設定されて登録されている文書（登録文書）を処理するための登録文書処理方法において、処理に応じたコンテンツを取り出すためのコンテンツバウンダリ条件が入力されると、そのコンテンツバウンダリ条件を受け付け、そのコンテンツバウンダリ条件に基づいて、前記登録文書に対して設定されているコンテンツバウンダリ情報を出力し、このコンテンツバウンダリ情報に対応するコンテンツバウンダリ位置を設定する処理を登録文書処理に含むものである。

【００１６】そして、前記コンテンツバウンダリ条件に基づいて、前記登録文書に対して設定されたコンテンツバウンダリ情報を出力する処理は、入力されたコンテンツバウンダリ条件と、前記登録文書に対して設定されたそれぞれのコンテンツバウンダリ情報とを比較し、前記コンテンツバウンダリ条件に適合するコンテンツバウンダリ情報を得て、このコンテンツバウンダリ情報に基づくコンテンツバウンダリ位置を設定するようにしている。

【００１７】また、本発明の請求項６に記載された登録文書処理方法の発明は、文書内容のまとまりの境界を示すコンテンツバウンダリ情報が設定されて登録された文書（登録文書）を処理するための登録文書処理方法において、処理に応じたコンテンツを取り出すために必要なコンテンツバウンダリ条件知識を格納したコンテンツバウンダリ条件知識ファイルを有し、このコンテンツバウンダリ条件知識ファイル内容に基づいて、前記登録文書に対して設定されたコンテンツバウンダリ情報を処理に応じて出力し、そのコンテンツバウンダリ情報に対応するコンテンツバウンダリ位置を設定する処理を登録文書処理に含むことを特徴としている。

【００１８】また、請求項７に記載の登録文書を処理するための処理プログラムを記憶した記憶媒体は、文書内容のまとまりの境界を示すコンテンツバウンダリ情報が設定されて登録されている文書（登録文書）を処理するための処理プログラムを記憶した記憶媒体であって、その処理プログラムは、処理に応じたコンテンツを取り出すためのコンテンツバウンダリ条件が入力されると、そのコンテンツバウンダリ条件を受け付け、そのコンテンツバウンダリ条件に基づいて、前記登録文書に対して設定されたコンテンツバウンダリ情報を出力し、このコンテンツバウンダリ情報に対応するコンテンツバウンダリ位置を設定する処理を含むものである。

【００１９】そして、前記コンテンツバウンダリ条件に基づいて、前記登録文書に対して設定されたコンテンツバウンダリ情報を出力する処理は、入力されたコンテンツバウンダリ条件と、前記登録文書に設定されたコンテンツバウンダリ情報とを比較し、前記コンテンツバウンダリ条件に適合するコンテンツバウンダリ情報を得て、このコンテンツバウンダリ情報に基づくコンテンツバウンダリ位置を設定するようにしている。

【００２０】また、本発明の請求項９記載の登録文書を処理するための処理プログラムを記憶した記憶媒体は、文書内容のまとまりの境界を示すコンテンツバウンダリ情報が設定されて登録されている文書（登録文書）を処理するための処理プログラムを記憶した記憶媒体であって、その処理プログラムは、登録文書に対する処理に応じて、処理に応じたコンテンツを取り出すために必要なコンテンツバウンダリ条件をコンテンツバウンダリ知識ファイルから取り出し、そのコンテンツバウンダリ条件に基づいて、前記登録文書に対して設定されたコンテンツバウンダリ情報を出力し、そのコンテンツバウンダリ情報に対応するコンテンツバウンダリ位置を設定する処理を登録文書処理に含むことを特徴としている。

【００２１】本発明は、コンテンツバウンダリ情報が設定されて登録された文書を用いて、検索や文書間の差分を取るといった様々な処理を行うものであり、これらの様々な処理を行う際、既に設定されているコンテンツバウンダリ情報を様々な処理に応じて選択することで、コンテンツバウンダリ位置を任意に設定することができる。これにより、処理に応じた適切な大きさのコンテンツを抽出することが可能となる。

【００２２】これを実現するための１つの手段として、ユーザがコンテンツバウンダリ条件を入力することにより、そのコンテンツバウンダリ条件に基づいて、前記登録文書に対して設定されているコンテンツバウンダリ情報の中から必要なコンテンツバウンダリ情報を選択し、このコンテンツバウンダリ情報に対応するコンテンツバウンダリ位置を設定する処理を行う。これは、ユーザがコンテンツバウンダリ条件を明示的に入力することによって、処理に必要なコンテンツを取り出すものであり、ユーザの意図を的確に反映したコンテンツを取り出すことができる。

【００２３】また、他の手段として、処理に応じたコン

テンツを取り出すために必要なコンテンツバウンダリ条件知識を持ち、このコンテンツバウンダリ条件知識に基づいて、前記登録文書に対して設定されたコンテンツバウンダリ情報の中から、処理に応じて選択して出力し、そのコンテンツバウンダリ情報に対応するコンテンツバウンダリ位置を設定することも可能である。これによれば、処理内容に応じて自動的に、処理に応じたコンテンツ条件が設定され、そのコンテンツ条件に基づいてコンテンツバウンダリ位置が設定されるので、ユーザが殆ど手を加えることなく、自動的に、処理に最適なコンテンツの抽出を行うことができる。

【００２４】このように、本発明は、登録文書に細かく設定されたコンテンツバウンダリ情報の中から、コンテンツバウンダリ情報を取捨選択することができるので、登録文書を用いて、文書間の差分を取ったり、検索したりする処理を行う際に、処理に必要なコンテンツを取り出す処理がきわめて簡単に行える。つまり、登録された文書を処理する段階において、予め設定されているコンテンツバウンダリ情報の中から、コンテンツを取り出すためのコンテンツバウンダリ情報を処理の種類に応じて任意に決めることができる。

【００２５】

【発明の実施の形態】以下、本発明に実施の形態について説明する。本発明は、登録された文書（登録文書）に対し、検索処理や文書間の差分を取るなど何らかの処理を施す際に文書を文章のまとまりを１つの単位として、そのまとまり（コンテンツ）に分割して取り出すための装置および方法に係わる発明であるが、まず、文書を登録する段階で、登録しようとする文書（以下、登録対象文書という）からコンテンツバウンダリ情報を取得して、そのコンテンツバウンダリ情報を保存する処理についてを説明し、その後で、その登録文書に対して文書を所定のコンテンツにて抽出する処理について説明する。

【００２６】図１は文書を登録する際に行われるコンテンツバウンダリ情報取得についての文書登録装置（以下、第１の文書登録装置という）の構成図である。この発明でいう、文書登録装置というのは、文書の作成、編集、保存などが可能な装置であり、たとえば、パーソナルコンピュータなどもその一例である。

【００２７】この第１の文書登録装置は、コンテンツバウンダリ入力手段１、コンテンツバウンダリ情報記憶手段２を少なくとも備えた構成となっている。

【００２８】このような構成において、登録対象文書３に対して、ユーザが、コンテンツバウンダリ入力手段１からコンテンツバウンダリの位置を指定すると、その位置に対応するコンテンツバウンダリ情報がコンテンツバウンダリ情報記憶手段２に記憶される。このコンテンツバウンダリ情報記憶手段２の記憶内容は、コンテンツバウンダリ情報ファイル４として保存することもできる。そして、そのコンテンツバウンダリ情報は、文書の中に通常は表示されないデータの形式で埋め込まれて保存されてもよいし、また、コンテンツ情報のみを文書データとは別のデータ（たとえば、コンテンツバウンダリ情報テーブル）として保存されるようにしてもよい。

【００２９】このように、第１の文書登録装置では、ユーザが登録対象文書３に対して、明示的にコンテンツバウンダリの位置を決めるものであり、具体的には次のようにして行う。

【００３０】たとえば、ディスプレイ画面１０上に映し出されている登録対象文書３が図２のような内容であったとする。このような文書内容に対して、ユーザは、その文書内容を見て、マウスなどにより、明示的にコンテンツバウンダリの位置を指定して行く。図２において、矢印はマウスカーソルｃを示しており、このマウスカーソルｃをユーザの意図する部分に位置させ、その位置でマウスをクリックすることによりコンテンツバウンダリ位置が設定される。図２では設定されたコンテンツバウンダリ位置をｂ１，ｂ２，ｂ３で示している。なお、このようなコンテンツバウンダリ位置を設定する場合は、システムのアプリケーションをコンテンツバウンダリ設定モードに設定して行う。

【００３１】また、この図２の例では、コンテンツバウンダリとする部分にマウスカーソルを位置させてクリックすることでコンテンツバウンダリ位置を指定するようにしたが、これに限らず、たとえば、コンテンツバウンダリで仕切られる文書内容（コンテンツという）の先頭にマウスカーソルを位置させて、そのコンテンツの終わりまでマウスカーソルをドラッグさせることによって、コンテンツバウンダリ位置を指定することも可能である。

【００３２】このようにして、コンテンツバウンダリ位置の指定がなされるが、そのコンテンツバウンダリ位置に対応するコンテンツバウンダリ情報は、前述したように、文書とは別のデータとして保存してもよく、文書の中に通常は表示されないデータの形式で埋め込んで保存してもよい。

【００３３】図３は図２で指定されたコンテンツバウンダリ情報を、文書の中に通常は表示されないデータの形式で埋め込んだ例を示すものである。図３の例では、ＨＴＭＬ（Hyper Text Markup Language）のコメントタグを用いて、文書の中に埋め込んだ例である。

【００３４】図３において、＜！－ＣＢ１・・・－＞が示される内容がコンテンツバウンダリ情報である。この＜！－ＣＢ１・・・－＞は、「＜！－」がコメントの開始を表し、「－＞」がコメントの終了を表している。

【００３５】一例として、＜！－ＣＢ１　ファイル端　１４０　１－＞というように記述されたコメントタグが有るとすると、その中のコンテンツバウンダリ情報として、「ＣＢ１」はコンテンツバウンダリの識別番号、「ファイル端」はコンテンツバウンダリの種類、「１４

０」は、そのコンテンツバウンダリの種類（この場合「ファイル端」）の反対側のコンテンツバウンダリの識別番号であり、そのコンテンツバウンダリ種類により仕切られるコンテンツの大きさをも表している。また、「１」はネストレベルを表している。なお、このようなコンテンツバウンダリ情報の詳細については後に説明する。

【００３６】また、前述の＜！－ＣＢ１　ファイル端　１４０　１－＞で示されるコメントタグには、コンテンツバウンダリの位置を表す内容は存在しないが、そのコンテンツバウンダリの位置は、このようなコメントタグの存在する位置であり、これによってコンテンツバウンダリの位置がわかるのでその位置情報は、特に、記述する必要はない。

【００３７】以上が第１の文書登録装置についての説明である。この第１の文書登録装置におけるコンテンツバウンダリ情報の抽出処理は、処理対象文書に対し、ユーザが明示的にコンテンツバウンダリ位置を指示することにより行われるものであり、ユーザが行うべき操作がやや面倒であるが、ユーザの意図する場所に確実にコンテンツバウンダリを設定することができるのが大きな特徴である。

【００３８】図４は登録対象文書を登録する際に行われるコンテンツバウンダリ情報取得を行うための第２の文書登録装置の構成図である。

【００３９】この第２の文書登録装置は、図１で示した第１の文書登録装置で示されたコンテンツバウンダリ情報記憶手段２、コンテンツバウンダリ情報ファイル４を備え、加えて、コンテンツバウンダリ条件入力手段５とコンテンツバウンダリ解析手段６を備え、さらに、必要に応じて、コンテンツバウンダリ情報表示手段７を備えた構成となっている。

【００４０】この第２の文書登録装置では、第１の文書登録装置のように、登録対象文書３に対して、ユーザが、コンテンツバウンダリの位置を明示的に指定するのではなく、コンテンツバウンダリ条件を入力することで、そのコンテンツバウンダリ条件に基づいて、自動的にコンテンツバウンダリを設定する。

【００４１】すなわち、コンテンツバウンダリ条件入力手段５により、ユーザがコンテンツバウンダリ条件の入力を行うと、コンテンツバウンダリ解析手段６により、ユーザによって設定されたコンテンツバウンダリ条件を解析する。ここでのコンテンツバウンダリ条件というのは、たとえば、段落、空行、改行、リスト、リスト項目、表など文書を１つのまとまりとして抽出できる部分である。このようなコンテンツ条件をユーザが入力すると、コンテンツバウンダリ解析手段６が処理対象文書をたとえばスキャンして、入力されたコンテンツバウンダリ条件に基づいて、コンテンツバウンダリ候補の位置や種類を抽出して、コンテンツバウンダリ情報として出力する。

【００４２】このコンテンツバウンダリ解析手段６によって得られたコンテンツバウンダリ情報は、コンテンツバウンダリ情報記憶手段２に記憶される。このとき、前述の第１の文書登録装置と同様に、このコンテンツバウンダリ情報記憶手段２の記憶内容は、コンテンツバウンダリ情報ファイル４として保存することもできる。そして、そのコンテンツバウンダリ情報は、文書の中に通常は表示されないデータの形式で埋め込まれて保存されてもよいし、また、コンテンツ情報のみを文書データとは別のデータ（たとえば、コンテンツバウンダリ情報テーブル）として保存されるようにしてもよい。

【００４３】また、コンテンツバウンダリ解析手段６によって得られたコンテンツバウンダリ情報は、コンテンツバウンダリ情報表示手段７に表示させ、ユーザの設定したコンテンツバウンダリ条件に対してどのようなコンテンツ情報が作成されたかをユーザに知らせるようにすることもできる。

【００４４】以下に具体例を参照しながらさらに説明する。

【００４５】図５に示すように、たとえば、ディスプレイ画面１０に登録対象文書３などの表示を行うための表示エリア１０ａと、コンテンツバウンダリ条件入力手段としての表示エリア１０ｂとを設定し、表示エリア１０ａには登録対象文書３を表示し、表示エリア１０ｂにはコンテンツバウンダリ条件を表示する。ここでは、コンテンツバウンダリ条件として、ファイル端、リスト、リスト項目、空行、改行、表などが示されている。

【００４６】そして、ユーザがたとえば、コンテンツバウンダリ条件として「リスト」を選択したとする（図５において、選択されたコンテンツバウンダリ条件は黒丸で示されている）。これにより、コンテンツバウンダリ解析手段６は、ユーザの設定したコンテンツバウンダリ条件に基づいて、登録対象文書３内のコンテンツバウンダリとなりうるコンテンツバウンダリ候補の位置および種類の解析を行い、その結果をコンテンツバウンダリ情報として出力する。

【００４７】そして、コンテンツバウンダリ解析手段６により得られるコンテンツバウンダリ情報に基づいて、コンテンツバウンダリ位置をディスプレイ画面１０の表示エリア１０ａ上で表示するとともに、前述の第１の文書登録装置で説明したように、コンテンツバウンダリ解析手段６により得られるコンテンツバウンダリ情報をコンテンツバウンダリ情報記憶手段２に記憶させる。

【００４８】ユーザはディスプレイ画面１０の表示エリア１０ａに表示された内容を見て、自分の意図したコンテンツバウンダリ位置が適正に反映されているか否かを判断し、修正したい箇所があればそれを指示することもできる。

【００４９】なお、コンテンツバウンダリ条件は、図５

で示した例のように、予め表示されている幾つかの条件から選択するというのではなく、ユーザがコンテンツバウンダリ条件や、コンテンツバウンダリとなりうるパターンなどを入力するようにしてもよい。このコンテンツバウンダリとなりうるパターンというのは、たとえば、文書中に、規則性のある字句や記号が繰り返し現れるような場合、その字句や記号をコンテンツバウンダリとして入力することもできる。たとえば、具体例として、「１日」という項目があって、その「１日」という項目のあとに、あるまとまった文章が存在し、行を変えて、「２日」という項目があって、その「２日」という項目についてのあるまとまった文章が存在するというように、規則性のある字句や記号が繰り返し現れるような文書内容があるとする。このような例では、「数字＋日」といったパターンをコンテンツバウンダリ条件として設定することができる。

【００５０】以上説明した第２の文書登録装置では、ユーザがコンテンツバウンダリ条件を入力するだけで、あとは、入力されたコンテンツバウンダリ条件に基づいて、自動的に処理対象文書に対するコンテンツバウンダリ情報を得ることができる。なお、この第２の文書登録装置は、第１の文書登録装置と同様に、ユーザが処理対象文書中に明示的にコンテンツバウンダリを指示することも可能である。

【００５１】図６は登録対象文書を登録する際に行われるコンテンツバウンダリ情報取得を行うための第３の文書登録装置の構成図である。

【００５２】この第３の文書登録装置は、図４の第２の文書登録装置で示されたコンテンツバウンダリ情報記憶手段２とコンテンツバウンダリ解析手段６、コンテンツバウンダリ情報ファイル４を備え、加えて、コンテンツバウンダリ知識が格納されたコンテンツバウンダリ知識ファイル８を備えた構成となっている。

【００５３】この第３の文書登録装置では、登録対象文書３に対し、コンテンツバウンダリ知識ファイル８を用いて、コンテンツバウンダリとなりうる部分をすべて自動的に検出し、それをコンテンツバウンダリ情報としてコンテンツバウンダリ情報記憶手段２に記憶させるものである。

【００５４】前述のコンテンツバウンダリ知識ファイル８には、コンテンツバウンダリとなりうる各種の条件が予め記述されている。このコンテンツバウンダリとなりうる各種の条件とういうのは、たとえば、句点、改行、空行、大文字見出し、ＨＴＭＬにおけるタグなどである。

【００５５】図７は第３の文書登録装置の処理手順を説明するフローチャートであり、まず、登録対象文書を入力して（ステップｓ１）、データの読み込みを行い、文書末であるか否かを判定し（ステップｓ２）、文書末であれば終了し、文書末でなければステップｓ３に進む。ステップｓ３では、登録対象文書にコンテンツバウンダリ候補が存在するか否かを判定し、存在しなければ、ステップｓ２に戻り、コンテンツバウンダリ候補が有れば、そのコンテンツバウンダリ候補に対してコンテンツバウンダリ情報を設定し、そのコンテンツバウンダリ情報をコンテンツバウンダリ情報記憶手段に記憶させる（ステップｓ４）。

【００５６】この図７のフローチャートで示される処理は、主に、コンテンツバウンダリ解析手段６が行う処理であり、読み込んだ登録対象文書と、コンテンツバウンダリ知識ファイルに記述されているコンテンツバウンダリ条件とを比較し、登録対象文書内にコンテンツバウンダリ知識ファイルに記述されているコンテンツバウンダリ条件に一致する部分が存在すると、コンテンツバウンダリ候補を検出したとして、その部分に対応するコンテンツバウンダリ情報を得て、そのコンテンツバウンダリ情報をコンテンツバウンダリ記憶手段２に記憶させる。

【００５７】そして、コンテンツバウンダリ記憶手段２では、受け取ったコンテンツバウンダリ情報を新たなコンテンツバウンダリ情報として格納する。このとき、前述の第１、第２の文書登録装置と同様に、コンテンツバウンダリ情報記憶手段２の記憶内容は、コンテンツバウンダリ情報ファイル４として保存することもできる。そして、そのコンテンツバウンダリ情報は、文書の中に通常は表示されないデータの形式で埋め込まれて保存されてもよいし、また、コンテンツバウンダリ情報のみを文書データとは別のデータ（たとえば、コンテンツバウンダリ情報テーブル）として保存されるようにしてもよい。

【００５８】以上説明した第３の文書登録装置では、登録対象文書に対し、コンテンツバウンダリ知識ファイル８の内容に基づいて、自動的にコンテンツバウンダリ情報を得ることができ、ユーザがコンテンツバウンダリを明示的に指示したり、コンテンツバウンダリ条件を入力したりする操作が不要となる。

【００５９】ところで、これまで説明した第１～第３の文書登録装置にて求められるコンテンツバウンダリ情報は、たとえば、図８に示すようなコンテンツバウンダリ情報テーブルとして表すことができる。以下、このコンテンツバウンダリ情報について図８のコンテンツバウンダリ情報テーブルを参照しながら説明する。

【００６０】図８で示されるコンテンツバウンダリ情報テーブルは、そのテーブルの１つの行で示される内容が１つのコンテンツバウンダリ情報であり、たとえば、第１行目の内容、つまり、コンテンツバウンダリ識別番号「１」、コンテンツバウンダリ位置「０」、コンテンツバウンダリの種類「ファイル端」、対応するコンテンツバウンダリ「１４０」、ネストレベル「１」は、図３において、ＨＴＭＬのコメントタグの一例として示した＜！－ＣＢ１　ファイル端　１４０　１－＞に対応するコ

ンテンツバウンダリ情報である。

【００６１】このようなコンテンツバウンダリ情報において、コンテンツバウンダリ識別番号は、その文書のコンテンツバウンダリとなりうる部分に付された番号である。

【００６２】また、コンテンツバウンダリ位置は、文書データの先頭からの文字数を表し、コンテンツバウンダリ位置が「０」というのは、文書ファイルの先頭を表している。また、コンテンツバウンダリの種類は、コンテンツバウンダリが文書のどのような部分かを表すもので、コンテンツバウンダリの種類が「ファイル端」というのは、コンテンツバウンダリがその文書の端であることを表している。

【００６３】そして、対応するコンテンツバウンダリというのは、コンテンツバウンダリの種類で指定されるコンテンツバウンダリの反対側に位置するコンテンツバウンダリの位置を、コンテンツバウンダリ識別番号で表すものである。

【００６４】たとえば、コンテンツバウンダリの種類が「ファイル端」で対応するコンテンツバウンダリが「１４０」であるとすると、ファイルの先頭の反対側のコンテンツバウンダリ位置、つまり、ファイルの終わりの位置が、コンテンツバウンダリ識別番号「１４０」であることを表している。

【００６５】また、図3のような文書内容において、ＨＴＭＬのコメントタグが、＜！－ＣＢ１５　リスト項目　１５　３－＞となっている場合は、図８で示されるコンテンツバウンダリ情報テーブルからわかるように、コンテンツバウンダリ識別番号は「１５」であり、コンテンツバウンダリ位置の「５０」は、文書のファイル先頭からの文字数が５１文字目（先頭が0から始まっているので、「５０」は５１文字目となる）を表している。

【００６６】また、コンテンツバウンダリの種類が「リスト項目」というのは、リストとして記述された幾つかの項目のうちの１つの項目であることを表している。そして、対応するコンテンツバウンダリ「１５」は、この場合、そのリスト項目自体を１つのコンテンツとすることを意味している。

【００６７】また、ネストレベルを示す数値は、このコンテンツバウンダリ情報テーブルで表されるように、最も大きなコンテンツをその文書ファイル全体としたとき、その文書ファイル全体のネストレベルを「１」とし、その中に、たとえば、リストという内容が1つのコンテンツとして存在した場合、そのリストによるコンテンツは、その文書ファイル全体で表されるコンテンツの中に含まれるので、ネストレベルを「２」とし、そのリストの中に存在するリスト項目は、ネストレベルを「３」とするというように、あるコンテンツの中に含まれるコンテンツ、さらにそのコンテンツの中に含まれるコンテンツというように、包含される度合いが高いほどネストレベルを表す数値が大きくなる。

【００６８】また、図８のコンテンツバウンダリ情報テーブルにおいて、たとえば、コンテンツバウンダリ識別番号「３」のコンテンツバウンダリは、そのコンテンツバウンダリ位置が「３０」であり、コンテンツバウンダリの種類が「句点」で、対応するコンテンツバウンダリが「２」、ネストレベルが「２」となっている。これは、対応するコンテンツバウンダリが「２」であることから、この場合、「句点」でコンテンツを仕切ると、ファイルの１１文字目から３１文字目（先頭が0から始まっているので、「１０」は１１文字目、「３０」は３１文字目となる）までを１つのコンテンツとするということであり、そのネストレベルは、ファイル全体を１つのコンテンツとして考えたとき、そのコンテンツ内に含まれるため、ネストレベルが「２」となっている。

【００６９】以上のようにして、処理対象文書中のコンテンツバウンダリ情報が作成され、そのコンテンツバウンダリ情報がコンテンツバウンダリ情報記憶手段２に記憶される。この図８に示すコンテンツバウンダリ情報テーブルにおいては、そのコンテンツバウンダリ情報テーブルにおける１つの行がそれぞれのコンテンツバウンダリ情報を示している。

【００７０】なお、このようなコンテンツバウンダリ情報において、バウンダリの種類は、コード化してもよい。たとえば、「ファイル端」は「１」、「句点」は「２」、「リスト」は「３」というようにコード化して、そのコードデータを記憶するようにしてもよい。また、バウンダリ位置は文字数でなくても、バイト数でもよく、また行数でもよい。

【００７１】以上が文書を登録する際に、コンテンツバウンダリ情報を得て文書の登録を行う処理である。ところで、これまでの説明では、登録対象文書そのものを登録する処理についての説明はなされていないが、この登録対象文書は所定の登録手段に登録されることはいうまでもない。この登録は、前述したように、コンテンツバウンダリ情報とは別のデータとして登録されてもよく、あるいは、コンテンツバウンダリ情報が埋め込まれた状態で登録されてもよい。

【００７２】このようにして、コンテンツバウンダリ情報を持って登録された文書に対し、検索処理や複数の文書間の差分をとるなど何らかの処理を行おうとする際、すでに設定されたコンテンツバウンダリ情報を用いることで、これらの様々な処理に対応した適切なコンテンツを取り出すことができ、それぞれの処理を円滑にかつ容易に行うことができる。なお、処理の種類などによっては、その処理の種類に応じた適切な大きさのコンテンツを抽出する必要がある。このように、登録時においては、コンテンツそのものを決めるのではなく、コンテンツを取り出すためのコンテンツバウンダリ情報を細かく求めているので、登録された文書を処理する段階で、コ

ンテンツバウンダリ情報を取捨選択することが可能であり、それによって、取り出すコンテンツの大きさを柔軟に設定することができ、処理に対応したコンテンツを取り出すことを可能としている。

【００７３】なお、これまで説明した第１～第３の文書登録装置では、登録対象文書のコンテンツバウンダリ情報を得て、そのコンテンツバウンダリ情報をコンテンツバウンダリ情報テーブルとして保存したり、コンテンツ情報を文書中に埋め込んで保存したりすることを可能としている。

【００７４】したがって、このように登録された文書（登録文書）は、コンテンツ解析が行われているので、その登録文書に対して検索処理を行ったり、文書間の差分を取ったりする処理を行おうとする際、すでに設定されたコンテンツバウンダリ情報を用いることで、様々な処理に対応できるが、処理の内容などによっては、処理内容に応じた適切なコンテンツバウンダリ位置を設定する必要が出てくる場合もある。これに対処するために、登録文書に対して、以下のようなコンテンツ抽出処理を行う。

【００７５】図９は前述したような登録文書に対し適切な大きさのコンテンツを抽出する処理を行うための登録文書処理装置（以下、第１の登録文書処理装置という）の構成図である。

【００７６】この第１の登録文書処理装置は、コンテンツバウンダリ条件入力手段１１、コンテンツバウンダリ解析手段１２、コンテンツバウンダリ出力手段１３、コンテンツバウンダリ情報記憶手段１４、コンテンツバウンダリ情報ファイル１５、コンテンツバウンダリ情報表示手段１６を備えている。

【００７７】そして、コンテンツバウンダリ解析手段１２と、コンテンツバウンダリ出力手段１３とによって、コンテンツバウンダリ設定手段２０を構成している。このコンテンツバウンダリ設定手段２０は、コンテンツバウンダリ条件が入力されると、そのコンテンツバウンダリ条件に基づいて、前記文書登録時に得られたコンテンツバウンダリ情報を得て、このコンテンツバウンダリ情報によりコンテンツバウンダリ位置を示す情報を出力して、文書に対し、コンテンツバウンダリを設定するものである。

【００７８】なお、文書登録と登録文書処理を１つの同じシステムで行う場合は、コンテンツバウンダリ条件入力手段１１、コンテンツバウンダリ解析手段１２、コンテンツバウンダリ情報記憶手段１４、コンテンツバウンダリ情報ファイル１５、コンテンツバウンダリ情報表示手段１６などは、前述した文書登録装置（たとえば、第２の文書登録装置）で示したコンテンツバウンダリ条件入力手段５、コンテンツバウンダリ解析手段６、コンテンツバウンダリ情報記憶手段２、コンテンツバウンダリ情報ファイル４、コンテンツバウンダリ情報表示手段７と共用することができるが、ここでは、説明の都合上、これらを前述の文書登録装置とは別な符号を付して説明する。

【００７９】さらに、これらの構成要素の他に、コンテンツバウンダリ設定手段２０により設定されたコンテンツバウンダリ位置によって、文書をコンテンツに分割してそのコンテンツを抽出する手段なども実際には設けられるが、図９ではこれらを省略している。

【００８０】ところで、コンテンツバウンダリ情報記憶手段１４には、前述の文書登録装置によって得られたコンテンツバウンダリ情報が格納されている。また、このコンテンツバウンダリ情報は、コンテンツバウンダリ情報ファイル１５にファイルとして格納されていてもよい。なお、そのコンテンツバウンダリ情報は、前述したように、文書データとは別のデータとして登録されていてもよく、あるいは、文書データの中に埋め込まれた状態で登録されていてもよい。

【００８１】このような構成において、前述のようにして登録された文書を用いて何らかの処理を行うとする場合、その処理に必要なコンテンツを取り出すためのコンテンツバウンダリ条件をコンテンツバウンダリ条件入力手段１１より、ユーザが入力する。

【００８２】まず、ユーザが、コンテンツバウンダリ条件を入力する。このコンテンツバウンダリ条件の入力は、明示的な条件（たとえば、改行、リストなど）を入力してもよいし、ユーザが文書中のコンテンツバウンダリ位置を指定し、その指定した位置に存在するコンテンツバウンダリ情報から、システムが最適なものを選択するようにしてもよい。さらに、「数字＋日」といったパターンをコンテンツバウンダリ条件とすることもできる。なお、入力されたコンテンツバウンダリ条件が文書登録時にコンテンツバウンダリ情報に含まれていない場合には、入力されたコンテンツバウンダリ条件を解析することで、それを新たに追加することも可能である。

【００８３】以下、この第１の文書処理装置の動作例について説明する。

【００８４】たとえば、図８のようなコンテンツバウンダリ情報テーブルが得られている場合、ユーザがコンテンツバウンダリ位置「４３」を指定すれば、「リスト」がコンテンツバンダリ条件となり、コンテンツバウンダリ位置の「４３」～「８９」までの内容を１つのコンテンツとする。

【００８５】図１０（ａ）は「改行」と「リスト項目」をコンテンツバウンダリ条件とした場合、図１０（ｂ）は「リスト項目」をコンテンツバウンダリ条件とした場合ののコンテンツバウンダリ候補位置をそれぞれ示すもので、図においてマークＭがコンテンツバウンダリ位置を示している。このように、ディスプレイ画面１０上の表示エリア１０ｂに表示されるコンテンツバウンダリ条件の選択の仕方によって表示エリア１０ａのコンテンツ

バウンダリ位置が変化する。

【００８６】コンテンツバウンダリ解析手段１２は、ユーザによって指定されたコンテンツバウンダリ条件に適合するコンテンツバウンダリ情報を、コンテンツバウンダリ情報記憶手段１４から抽出して出力する。

【００８７】たとえば、コンテンツバウンダリ情報が、図８で示されるようなコンテンツバウンダリ情報テーブルの形式で保存されている場合、ユーザの指定したコンテンツバウンダリ条件が「リスト項目」であるとすると、バウンダリ識別番号「１５」、「１７」が抽出されるが、このとき、文書の基本構造を作っている識別番号「１」と「１４０」で表されるファイル端、リスト構造を作っている識別番号「１３」と「２４」で表されるリストも抽出する。

【００８８】これにより、抽出されるコンテンツとしては、そのコンテンツをコンテンツバウンダリ位置で表すと、「０」～「４２」、つまり、ファイル端（ファイルの先頭）からリスト（リスト開始）の前までが１つのコンテンツとして抽出され、コンテンツバウンダリ位置「５０」～「５６」、つまり、リスト項目（最初のリスト項目）から２番目のリスト項目の前までが１つのコンテンツとして抽出され、コンテンツバウンダリ位置「５７」～「・・・」、つまり、２番目のリスト項目からそのリスト項目の終わりまで（この表では位置が記載されていないので、「・・・」で表す）が１つのコンテンツとして抽出され、コンテンツバウンダリ位置「・・・」～「８８」、つまり、あるリスト項目の終わりからリスト（リストの終わり）の前までが１つのコンテンツとして抽出されるというように、文書の内容のまとまりがそれぞれコンテンツとして抽出されることになる。

【００８９】図１１はコンテンツバウンダリ情報がコンテンツバウンダリ情報テーブル形式である場合の処理の手順を示すフローチャートである。この図１１に示す処理手順は、主に、コンテンツバウンダリ設定手段２０が行う処理であり、以下、このフローチャートを参照しながら説明する。

【００９０】まず、コンテンツバウンダリ情報テーブルの内容を読み込む（ステップｓ１１）。そして、ユーザによってコンテンツバウンダリ条件が入力されると、そのコンテンツバウンダリ条件を受け付ける（ステップｓ１２）。次に、読み込んだコンテンツバウンダリ情報のうち１つのコンテンツバウンダリ情報（コンテンツバウンダリ情報テーブルの１行分のコンテンツバウンダリ情報）を読み込む。このとき、読み込むべきコンテンツバウンダリ情報が存在するか否かを判定し（ステップｓ１３）、コンテンツバウンダリ情報が存在すれば、その読み込んだコンテンツバウンダリ情報が、ユーザからによって与えられたコンテンツバウンダリ条件に適合するか否かを判定する（ステップｓ１４）。そして、読み込んだコンテンツバウンダリ情報がコンテンツバウンダリ条件に適合すれば、そのコンテンツバウンダリ情報を出力する（ステップｓ１５）。

【００９１】次に、再びステップｓ１３に戻って、読み込んだコンテンツバウンダリ情報のうち次のコンテンツバウンダリ情報（コンテンツバウンダリ情報テーブルにおける次の１行分のコンテンツバウンダリ情報）を読み込み、読み込むべきコンテンツバウンダリ情報が存在するか否かを判定し（ステップｓ１３）、コンテンツバウンダリ情報が存在すれば、その読み込んだコンテンツバウンダリ情報がユーザによって与えられたコンテンツバウンダリ条件に適合するか否かを判定する（ステップｓ１４）。そして、読み込んだコンテンツバウンダリ情報がコンテンツバウンダリ条件に適合すれば、そのコンテンツバウンダリ情報を出力する（ステップｓ１５）。

【００９２】このような処理をコンテンツバウンダリ情報テーブルのすべてのコンテンツバウンダリ情報について行い、コンテンツバウンダリ情報テーブルのすべてのコンテンツバウンダリ情報について処理が終了すると、これまでの処理により選択されたコンテンツバウンダリ情報に対応するコンテンツバウンダリ位置を示すマークが付された文書をコンテンツバウンダリ情報表示手段１６に表示する（ステップｓ１６）。

【００９３】ここで、ユーザがその表示内容（たとえば、図１０）を見て、ユーザの意図する位置にコンテンツバウンダリが適切に付されているか否かを判定し（ステップｓ１７）、適切に付加されていれば、ＯＫとして処理を終了し、ユーザの意図しない位置にコンテンツバウンダリが付加されているような場合、あるいは、ユーザの意図する位置にコンテンツバンダリが付されていない場合は、コンテンツバウンダリ条件を変更するなどして、再度、ステップｓ１２以降の処理を行う。

【００９４】このように、ユーザがコンテンツバウンダリ条件を与えることによって、そのコンテンツバウンダリ条件に適合するコンテンツバウンダリ情報を選んで、そのコンテンツバウンダリ位置を文書中に設定することができ、しかも、そのコンテンツバウンダリ位置をコンテンツバウンダリ情報表示手段１６によってディスプレイ画面上に表示させることにより、ユーザは、その表示内容を見て、コンテンツバウンダリ位置が適切に設定されているか否かを判断することができ、適切でなければ、コンテンツバウンダリ条件を設定し直すなどして、再度、コンテンツバウンダリ位置の設定処理を行うことも可能となる。また、ディスプレイ画面上に表示された文書に対してマウスなどで直接、コンテンツバウンダリ位置を指示することもできる。

【００９５】以上の処理は、コンテンツバウンダリ情報が、コンテンツバウンダリ情報テーブル形式で保存されている場合の処理手順を示すフローチャートである。これに対して、コンテンツバウンダリ情報が、文書中にたとえばＨＴＭＬなどのコメントタグとして埋め込まれて

いる場合の処理は、図１２のような処理手順にて行われる。以下、図１２のフローチャートを参照しながらその処理手順について説明する。

【００９６】まず、ユーザによってコンテンツバウンダリ条件の入力が行われると、そのコンテンツバウンダリ条件を受け付ける（ステップｓ２１）。そして、文書データ（たとえば、ＨＴＭＬなどのコメントタグによるコンテンツバウンダリ情報が埋め込まれている文書データ）を読み込む。このとき、読み込むべき文書データが存在するか否かを判定し（ステップｓ２２）、文書データが存在すれば、その文書データ中に存在する最初のコメントタグで示されるコンテンツバウンダリ情報が、ユーザによって与えられたコンテンツバウンダリ条件に適合するか否かを判定する（ステップｓ２３）。そして、埋め込まれているコンテンツバウンダリ情報がコンテンツバウンダリ条件に適合すれば、そのコンテンツバウンダリ情報を出力する（ステップｓ２４）。

【００９７】次に、再びステップｓ２２に戻って、読み込むべき文書データが存在するか否かを判定し、文書データがあれば、その文書データ中に存在するコメントタグで示されるコンテンツバウンダリ情報のうち、２番目のコンテンツバウンダリ情報が、ユーザによって与えられたコンテンツバウンダリ条件に適合するか否かを判定する（ステップｓ２３）。そして、読み込んだコンテンツバウンダリ情報が入力されたコンテンツバウンダリ条件に適合すれば、そのコンテンツバウンダリ情報を出力する（ステップｓ２４）。

【００９８】このような処理を文書データの終わりまで行い、処理が終了すると、これまでの処理により選択されたコンテンツバウンダリ情報に対応するコンテンツバウンダリ位置を示すマークの付された文書をコンテンツバウンダリ情報表示手段１６によってディスプレイ画面上に表示する（ステップｓ２５）。

【００９９】ここで、ユーザがその表示内容（たとえば、図１０）を見て、ユーザの意図する位置にコンテンツバウンダリが適切に付されているか否かを判定し（ステップｓ２６）、ユーザの意図する位置にコンテンツバウンダリが適切に付加されていれば、ＯＫとして処理を終了し、ユーザの意図しない位置にコンテンツバウンダリが付加されているような場合、あるいは、ユーザの意図する位置にコンテンツバンダリが付されていない場合は、コンテンツバウンダリ条件を変更するなどして、再度、ステップｓ２１以降の処理を行う。

【０１００】この場合も前述したと同様に、ユーザがコンテンツバウンダリ条件を与えることによって、そのコンテンツバウンダリ条件に適合するコンテンツバウンダリ情報を選んで、そのコンテンツバウンダリ位置を文書中に設定することができる。また、そのコンテンツバウンダリ位置が設定された文書をコンテンツバウンダリ情報表示手段１６によってディスプレイ画面上に表示させることにより、ユーザは、その表示内容を見て、コンテンツバウンダリ位置が適切に設定されているか否かを判断することができ、適切でなければ、コンテンツバウンダリ条件を設定し直すなどして、再度、コンテンツバウンダリ位置の設定処理を行うことも可能となる。また、ディスプレイ画面上に表示された文書に対してマウスなどで直接、コンテンツバウンダリ位置を指示することもできる。

【０１０１】以上のような処理を行うことによって、処理対象文書に対し、コンテンツバウンダリ位置の設定が任意に行え、設定されたコンテンツバウンダリ位置によって、コンテンツを抽出することができる。これにより、検索処理や文書間の差分を取るなどといった様々な処理に応じた適切なコンテンツを取り出すことができる。

【０１０２】これら図１１および図１２で説明した２つの処理において、図１１の処理は、コンテンツバウンダリ情報テーブルにおけるコンテンツバウンダリ識別番号の数だけ処理を繰り返せばよく、一方、図１２の方は、文書データの文字数分の処理を繰り返す必要がある。たとえば、前述の図８のコンテンツバウンダリ情報テーブルの例で考えると、図１１の処理は、コンテンツバウンダリ情報の数は１４０個であるため、１４０回処理を繰り返せばよいが、図１２の方は、文字数が０～６４０８の６４０９個存在するため、６４０９回処理を繰り返す必要があるので、図１１の方が処理量が少なくて済む。

【０１０３】図１３は前述したような登録文書に対し適切な大きさのコンテンツを抽出する処理を行うための第２の登録文書処理装置の構成図である。

【０１０４】この第２の登録文書処理装置は、コンテンツバウンダリ出力手段１３とコンテンツバウンダリ情報検索手段１７で構成されるコンテンツバウンダリ設定手段２０と、コンテンツバウンダリ条件知識ファイル２１を備えた構成となっている。

【０１０５】そして、この第２の登録文書処理装置は、ユーザがコンテンツバウンダリ条件を設定するのではなく、処理の内容に応じて自動的にコンテンツバウンダリ条件の設定を可能としたものであり、それを実現するために、コンテンツバウンダリ条件知識ファイル２１を備えている。このように、この第２の登録文書処理装置では、コンテンツバウンダリ条件知識ファイル２１によって、自動的に、処理に最適なコンテンツ条件を得て、そのコンテンツ条件に基づいて、処理対象の文書（登録文書２２）に対してコンテンツバウンダリ位置を設定することが可能となるので、ユーザがコンテンツ条件を入力するためのコンテンツバウンダリ条件入力手段や、ユーザの入力したコンテンツバウンダリ条件に基づいたコンテンツバウンダリ情報を表示するためのコンテンツバウンダリ情報表示手段などは特に必要ではない。

【０１０６】コンテンツバウンダリ条件知識ファイル２

1は、処理内容に応じて最適なコンテンツバウンダリ条件が蓄えられているもので、たとえば、この第2の文書処理装置を情報検索システムに用いるものとすれば、情報検索に最適なコンテンツバウンダリ条件についての知識が蓄えられている。

【0107】なお、コンテンツバウンダリ条件知識ファイル21は、そのシステムが行う処理に合わせたコンテンツバウンダリ条件をのみを持たせることも可能であるが、様々な処理内容に適応できるように、様々な処理内容に応じた最適なコンテンツバウンダリ条件知識を蓄えておくことも勿論可能である。この場合、たとえば、ユーザが処理の種類などについての指令を、コンテンツバウンダリ情報検索手段17に与えることで、コンテンツバウンダリ情報検索手段17は、その指令に基づいて、処理に最適なコンテンツバウンダリ条件を、コンテンツバウンダリ条件知識ファイル21から得て、その条件に基づいて、登録文書22（コンテンツバウンダリ情報が埋め込まれている文書）を読み出して、与えられたコンテンツバウンダリ条件に対応するコンテンツバウンダリ情報を自動的に検索する。

【0108】このように、第2の登録文書処理装置は、システム側の持っているコンテンツバウンダリ条件知識によって、文書の中に埋め込まれているコンテンツバウンダリ情報の中から自動的に、コンテンツバウンダリ条件知識ファイル21が持っているコンテンツバウンダリ条件に適合するコンテンツバウンダリ情報を検索するので、処理対象文書に対し、より一層、自動化されたコンテンツバウンダリ位置設定が可能となる。

【0109】以上説明したように、本発明では、登録された文書を、たとえば、検索処理や文書間の差分を取るなど様々な処理に用いる場合、登録時に設定されたコンテンツバウンダリ情報を用いてコンテンツバウンダリ位置を決め、これにより、処理に必要なコンテンツの抽出を行うようにしているので、コンテンツの取り出しがきわめて簡単に行うことができる。たとえば、処理の内容に応じて、文書登録の際に設定された多数のコンテンツバウンダリ情報のうち、必要なコンテンツバウンダリ情報のみ選択することが可能であり、これにより、コンテンツバウンダリ位置を任意に決めることができ、処理の内容に応じて、ユーザの意図する大きさのコンテンツの抽出が可能となる。また、文書登録のときに設定されなかったコンテンツバウンダリ情報を、文書処理の時に新たに追加するということも比較的容易に行うことが可能となる。

【0110】なお、本発明は以上説明した各実施の形態に限定されるものではなく、本発明の要旨を逸脱しない範囲で種々変形実施可能となるものである。また、以上説明した本発明の登録文書の処理を行うための処理プログラムは、フロッピィディスク、光ディスク、ハードディスクなどの記録媒体に記録させておくことができ、本発明はその記録媒体をも含むものである。また、ネットワークから処理プログラムを得るようにしてもよい。

【0111】

【発明の効果】本発明では、コンテンツバウンダリ情報が設定された登録文書を、たとえば、検索に用いたり、文書間の差分を取るなど様々な処理に用いる場合、ユーザがコンテンツバウンダリ条件を入力することにより、そのコンテンツバウンダリ条件に基づいて、前記登録文書に対して設定されているコンテンツバウンダリ情報の中から必要なコンテンツバウンダリ情報を選択し、このコンテンツバウンダリ情報に対応するコンテンツバウンダリ位置を設定することができる。これは、ユーザがコンテンツバウンダリ条件を明示的に入力することによって、処理に必要なコンテンツを取り出すものであり、ユーザの意図を的確に反映したコンテンツを取り出すことができる。

【0112】また、コンテンツバウンダリ条件を入力することなく、自動的に、処理に最適なコンテンツの抽出を行うことができる。これは、処理に応じたコンテンツを取り出すために必要なコンテンツバウンダリ条件知識を持ち、このコンテンツバウンダリ条件知識に基づいて、前記登録文書に対して設定されたコンテンツバウンダリ情報の中から、処理に応じて選択して出力し、そのコンテンツバウンダリ情報に対応するコンテンツバウンダリ位置を設定するものであり、これによれば、処理内容に応じて自動的に、処理に応じたコンテンツ条件が設定され、そのコンテンツ条件に基づいてコンテンツバウンダリ位置が設定されるので、ユーザが殆ど手を加えることなく、自動的に、処理に最適なコンテンツの抽出を行うことができる。

【0113】このように、本発明では、登録時に得られたコンテンツバウンダリ情報をそのまますべて用いるのではなく、文書を処理する段階で、コンテンツバウンダリ情報を取捨選択することが可能であって、これによって、任意の位置にコンテンツバウンダリの設定が可能となり、取り出すコンテンツの大きさを柔軟に設定することができ、処理に対応したコンテンツを取り出すことができる。このように、コンテンツの大きさを柔軟に決めることができることから、検索処理や文書間の差分を取るなどの処理以外にも、文書を所定のコンテンツに分割して処理を行う必要のある様々な処理に対応することができる。

【図面の簡単な説明】

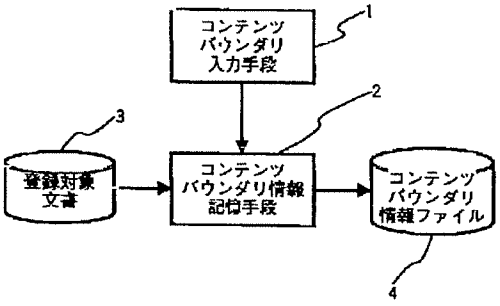
【図1】本発明の登録文書処理装置を実現するために必要な文書登録装置の第1の構成例（第1の文書登録装置）を説明するブロック図。

【図2】第1の文書登録装置におけるコンテンツバウンダリ位置指定についての一例を説明する図。

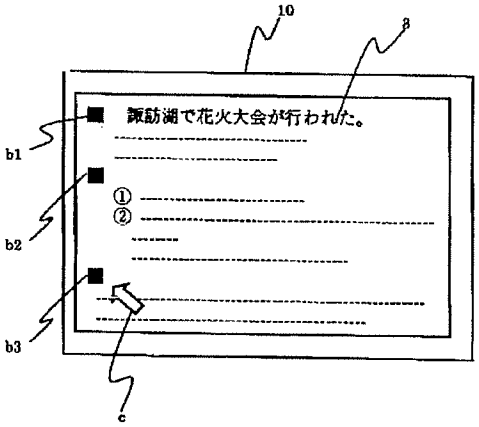
【図3】第1の文書登録装置におけるコンテンツバウンダリ情報を文書中に埋め込んだ例を示す図。

【図４】本発明の登録文書処理装置を実現するために必要な文書登録装置の第２の構成例（第２の文書登録装置）を説明するブロック図。
【図５】第２の文書登録装置におけるコンテンツバウンダリ条件入力を行う例を説明する図。
【図６】本発明の登録文書処理装置を実現するために必要な文書登録装置の第３の構成例（第３の文書登録装置）を説明するブロック図。
【図７】第３の文書登録装置におけるコンテンツバウンダリ情報を抽出する処理を説明するフローチャート。
【図８】第１～第３の文書登録装置において得られるコンテンツバウンダリ情報をコンテンツバウンダリ情報テーブルとして表した図。
【図９】本発明の実施の形態である第１の登録文書処理装置を説明するブロック図。
【図１０】本発明の実施の形態においてコンテンツバウンダリ条件の設定の仕方の違いよるコンテンツバウンダリ位置の変化を説明する図。
【図１１】第１の登録文書処理装置においてコンテンツバウンダリ情報がコンテンツバウンダリ情報テーブルである場合の処理手順を説明するフローチャート。
【図１２】第１の登録文書処理装置においてコンテンツバウンダリ情報が文書中に埋め込まれた場合の処理手順を説明するフローチャート。
【図１３】本発明の実施の形態である第２の登録文書処理装置を説明するブロック図。
【符号の説明】
１　コンテンツバウンダリ入力手段
２，１４　コンテンツバウンダリ情報記憶手段
３　登録対象文書
４，１５　コンテンツバウンダリ情報ファイル
５，１１　コンテンツバウンダリ条件入力手段
６，１２　コンテンツバウンダリ解析手段
７，１６　コンテンツバウンダリ情報表示手段
８　コンテンツバウンダリ知識ファイル
１３　コンテンツバウンダリ出力手段
１７　コンテンツバウンダリ情報検索手段
２０　コンテンツバウンダリ設定手段
２１　コンテンツバウンダリ条件知識ファイル
２２　登録文書

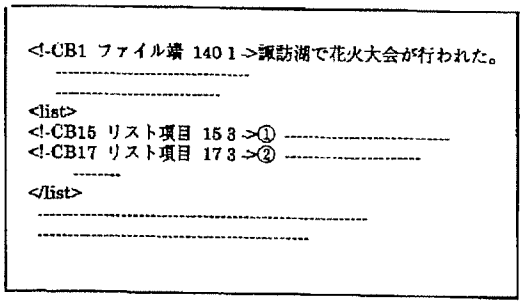
【図１】



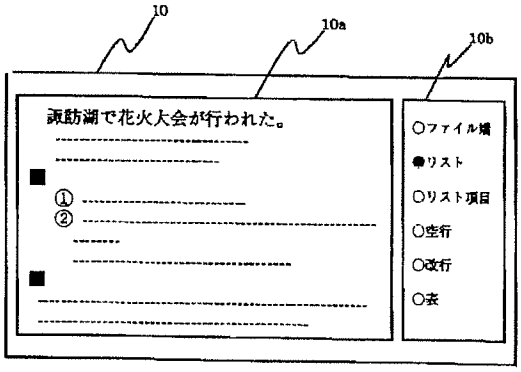
【図２】



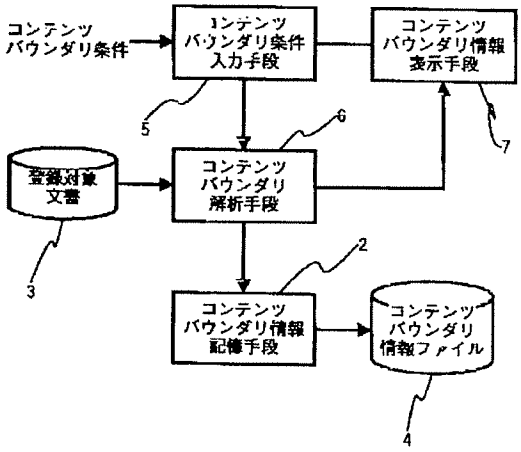
【図３】



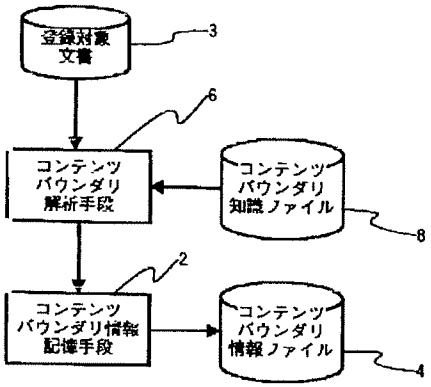
【図５】



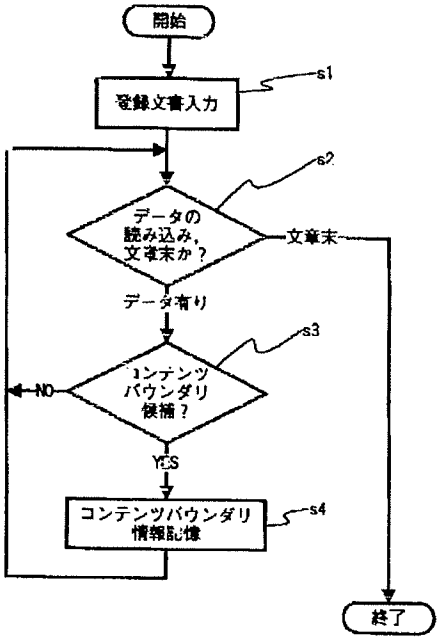
【図4】



【図6】



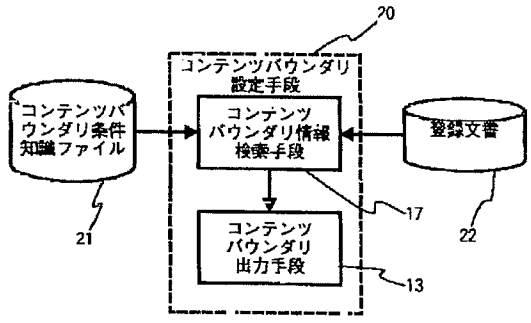
【図7】



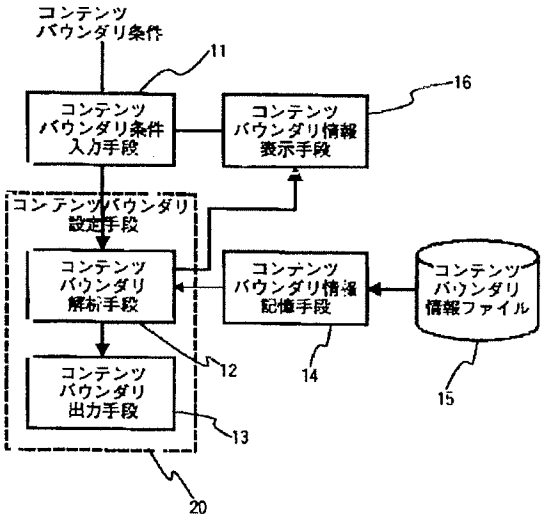
【図8】

コンテンツバウンダリ識別番号	コンテンツバウンダリ位置	コンテンツバウンダリの種類	対応するコンテンツバウンダリ	ネストレベル
1	0	ファイル端	140	1
2	10	句点	1	2
3	30	句点	2	2
4	31	改行	1	2
13	43	リスト	24	2
14	49	改行	13	2
15	50	リスト項目	15	3
16	56	改行	15	3
17	57	リスト項目	17	3
24	89	リスト	13	2
25	96	改行	24	2
140	6408	ファイル端	1	1

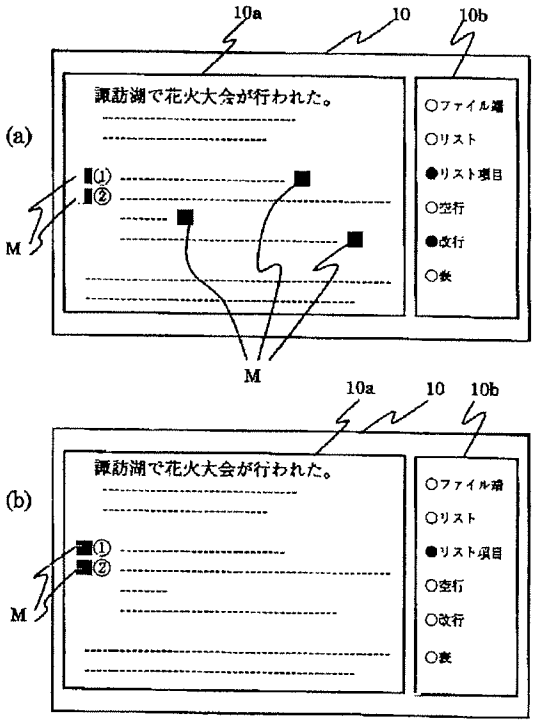
【図13】



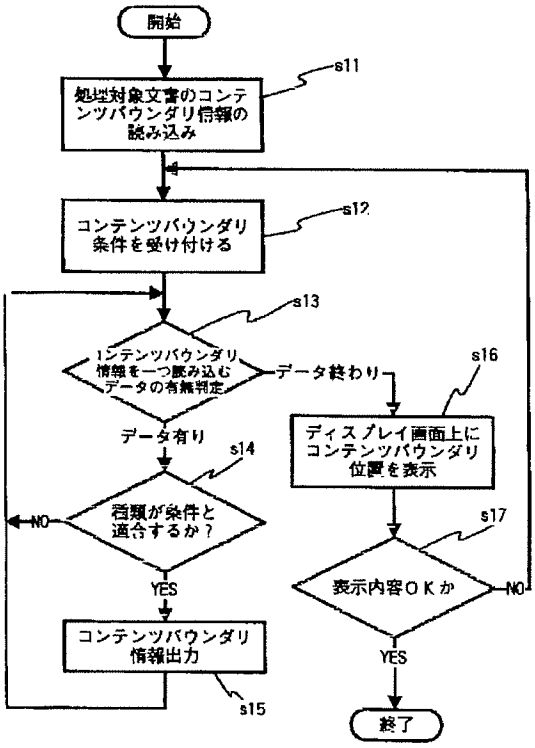
【図9】



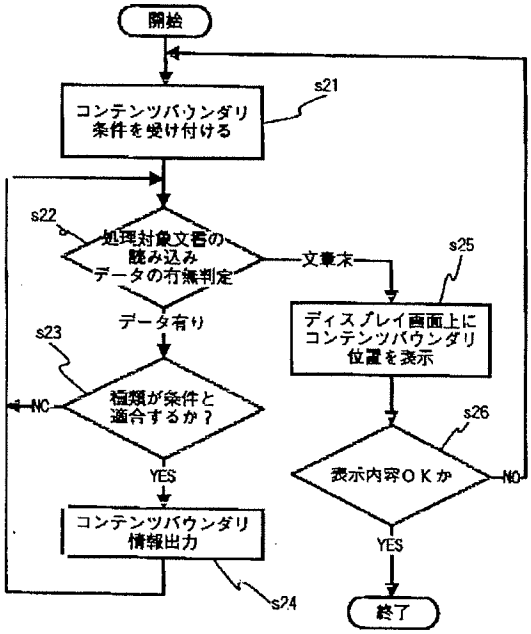
【図10】



【図11】



【図12】



フロントページの続き

(51)Int.Cl.[6]	識別記号	FI		
		G06F	15/401	310A
			15/403	340A